Branching and Inference in Population Genetics

Neil O'Connell*

March 28, 1994

Abstract

The probabilistic structure of the genealogy in branching processes is described; in particular, we present an analogue of Kingman's coalescent for near-critical branching processes. This result is applied to the problem of estimating the age of our most recent common ancestor using samples of mtDNA taken from contemporary humans. We also discuss more general issues concerning the use of models for making making inference about the past of a population.

1 Introduction

One of the most exciting detective stories in modern times began with Darwin's theory of evolution. Scientists have since been investigating the mysteries of human origin with remarkable success using natural clues such as fossils, bones and, more recently, molecules. The molecule in question is DNA, which is difficult to interpret but potentially the most informative clue of all: we simply have to learn how to read it. Given that we understand the mechanisms by which DNA evolves, we can attempt to make inference about

*Dublin Insitute for Advanced Studies, 10 Burlington Road, Dublin 4, Ireland. Research supported by NSF grants MCS90-01710 and DMS91-58583, and by grants from EOLAS and Mentec Computer Systems Ltd. This article will appear in the Proceedings of the IMA Workshop on Population Genetics, January 1994.

the past using DNA samples. In fact, geneticists have being doing this for some time.

One of the central questions people have considered is 'when and where (if at all) did the most recent common ancestor of all humans live?' In 1991, Vigilant *et al.* [23] claimed to have found molecular evidence for a recent African origin, using samples of mitochondrial DNA (mtDNA) from contemporary humans. Their estimate of the age of our most recent common ancestor, more affectionately referred to as 'Eve', is 200,000 years. It has since become a very controversial topic: there is a vast literature supporting variations of this hypothesis and an equally vast literature in opposition, finding fault in the methods used and quoting contradictory fossil evidence. Popular accounts can be found in [10, 21, 24]. This controversy is a symptom of the fact that here we have a difficult statistical problem.

One of the difficulties is that to make inference about the age and whereabouts of our most recent common ancestor it seems necessary to make assumptions about the genealogical dynamics of the population. The approach of Vigilant *et al.* [23] assumes that the true phylogeny relating individuals in a sample is the one which requires fewest mutations in order to explain the variety of DNA types observed: this method of reconstructing phylogenies is known as *parsimony*, and has been subjected to criticism on various grounds (see, for example, [5]). One of the problems with this approach is that often there are several, equally 'parsimonious', possible phylogenies, each leading to a different conclusion; another is the fact that there is no theoretical basis for parsimony. There is also the added difficulty of rooting the inferred trees.

Some authors (see, for example, [6]) have attempted to use maximum likelihood methods, but these are difficult to formulate properly and become very complicated when a large sample is used.

In this paper we argue that it is not necessary to construct a tree in order to make inferential statements about the *age* of Eve. We will present an alternative approach, where the genealogy is modelled via a branching process. We do not attempt to consider the whereabouts of Eve as this would require a spatial component in the model, which we haven't included.

The idea of adopting a statistical model for genealogical dynamics in order to make inference about the past has also been applied by Lundstrom, Tavaré and Ward [11] and Tavaré and Griffiths [20], where traditional (Wright-Fisher/Moran type) population genetics models are used. The application of branching processes in evolution is a more recent development,

and seems very promising; Jagers, Nerman and Taib [7, 9, 18, 19] have done a considerable amount of work on this topic. For general background on biological applications of branching processes, see [8].

The relationship between branching and traditional models for evolution is well known [17, 14, 3]. Qualitatively, both models display similar behaviour unless there is some kind of spatial structure, in which case the behaviour can be radically different. We will discuss these issues in §5.

The outline of the paper is as follows. In §2 we present some informal arguments justifying the use of a statistical model for population dynamics. In §3 we describe the probabilistic structure of genealogical trees in branching processes; in particular we present an analogue of Kingman's coalescent as an approximation for the family tree in a near-critical branching process. In §4 we apply these results to the problem of estimating the age of Eve. Finally, in §5, we discuss the relationship between branching and traditional models for evolution.

2 Using a model to estimate the age of Eve

Suppose we have a model that describes the statistical nature of the past evolution of the human species, and suppose one of the parameters of the model is taken to be the time back, T, to Eve. Suppose further that there exists a constant γ such that the expected time back to the most recent common ancestor of two randomly chosen individuals alive today is γT. Then, under the usual assumptions of neutral evolution, if δ denotes the mean rate of divergence along distinct lines of descent, the expected divergence between two randomly chosen individuals is $\delta\gamma T$. It follows that the expected mean pairwise divergence $\bar{d}$ in a random sample of individuals is also given by $\delta\gamma T$.

If δ and γ are known, this yields a straightforward moment estimate

$$\hat{T} := \frac{\bar{d}}{\delta\gamma}. \tag{1}$$

In practice the parameters δ and γ are *not* known; however, given reliable estimates, the estimator (1) can be approximated.

In the next section we show that for branching processes there is such a constant γ that depends on the mean growth rate of the population. This

fact is combined with (1) in §4 to provide a method for the simultaneous estimation of the mean population growth rate and T, the age of Eve.

3 The genealogy of branching processes

Let Z be a Markov branching process with mean lifetime 1, offspring distribution ν and let ξ be a realisation of ν. Fix $t > 0$, and for each $0 \leq s \leq t$, define $N_t(s)$ to be the number of individuals alive at time s with descendents alive at time t. The process N_t is called the *reduced branching process*, and can be thought of as the family tree relating the individuals alive at time t. It is also referred to as the *reduced family tree*. Note that N_t is also a Markov process.

When t is large, the only case where the genealogy is non-trivial is when the branching process is close to critical ($E\xi \approx 1$). To make this more precise, suppose $E\xi = 1 + \alpha/t$, for some $\alpha \in \mathbb{R}$. In [13] it is shown that when t is large and the time units are taken as t generations, the reduced process can be approximated by a linear pure birth process $\{N(r),\ 0 \leq r < 1\}$ with jump rate $b(\alpha, r)N(r)$ at time r, where

$$b(\alpha, r) = \begin{cases} \alpha\left(1 - e^{-\alpha}\right)^{-1}(1-r)^{-1} & \alpha \neq 0, \\ (1-r)^{-1} & \alpha = 0. \end{cases} \tag{2}$$

This result provides a complete probabilistic description of the family tree relating those individuals alive at time t; it is thus the branching process analogue of Kingman's coalescent. Note that, just as the coalescent can be used to speed up simulations, so can this result. It can also be used to describe the genealogical relationship between two randomly chosen individuals. If S_t denotes the time back to the most recent common ancestor of two individuals chosen randomly from the population at time t ($S_t = t$ if they have no common ancestor), then [13, Theorem 2.3] the laws of S_t/t conditional on $\{N_t(0) = x\}$ converge weakly, as $t \to \infty$, to a limiting law μ_x, say, on $(0, 1]$, defined by

$$\mu_x(0, r) = \frac{2q_r^x}{(x-1)!}\left\{(1-q_r)^{-x} - F(x-1, 1-q_r)\right\}, \tag{3}$$

for $0 \leq r \leq 1$, where

$$q_r = \frac{e^{-(1-r)\alpha} - e^{-\alpha}}{1 - e^{-\alpha}}, \tag{4}$$

and $F : \mathbb{Z}_+ \times (0,1) \to \mathbb{R}$ is defined by

$$F(n,y) = \frac{\partial^n}{\partial y^n}\left\{\frac{\log(1-y)}{y^2}\right\}. \tag{5}$$

It follows (applying bounded convergence) that

$$ES_t/t \to \gamma_x(\alpha), \tag{6}$$

where $\gamma_x(\alpha)$ denotes the mean of μ_x. In other words, if t is large,

$$ES_t \simeq \gamma_x(\alpha)t. \tag{7}$$

We remark that γ is increasing and $\gamma(\alpha) \nearrow 1$ as $\alpha \to \infty$.

In the supercritical case, when the process is not 'close' to critical, individuals are typically distantly related. For example, it follows from results of Bühler [1], Zubkov [25] and Durrett [4], that if $E\xi > 1$, $(S_t/t|\ Z(t) > 0) \to 1$ in probability as $t \to \infty$. This fact can be extrapolated (in some sense) from (2) by letting $\alpha \to \infty$. In the subcritical case ($E\xi < 1$) individuals typically have very recent ancestors: in this case, $(S_t/t|\ Z(t) > 0) \to 0$ in probability [4, 25]. Again this can be seen from (2) by letting $\alpha \to -\infty$. Essentially what we are doing here is describing the continuum of non-trivial possibilities in between, which arise when the process is close to critical.

4 Estimating the age of Eve

It is thought that mtDNA is inherited primarily from the mother. This assumption allows us to restrict our attention to single-sex populations, and so we are not forced to make questionable assumptions about the mating behaviour of people. According to the molecular clock hypothesis, substitutions occur randomly along lines of descent at a constant rate. Neutrality is assumed; that is, the occurence of substitutions along a particular line of descent is independent of the family tree structure and geographical location of individuals, and that substitutions along distinct lines occur independently of each other. The divergence rate is very small, so over the time period we are considering here (the post-Eve period) we can assume that each substitution produces a new type, that is, reverse substitutions do not occur. Thus, if the most recent common ancestor of two individuals died s million years ago,

the number of differences between their mtDNA types will be approximately Poisson with mean $2us$, where u is the substitution rate (in units of number of substitutions per million years). Now suppose two individuals are sampled randomly from the current population, and δ denotes the rate of divergence (in units of percentage divergence per million years). Note that if l denotes the sequence length, then $\delta = 2u/l$. If we have a model for the genealogical structure of the population, then the expected amount of divergence between the mtDNA sequences of the two individuals will be equal to the expected time back to the common ancestor of the two individuals (under our model, in units of millions of years), multiplied by the divergence rate, δ.

We will assume that the (effective) female population size follows a Markov branching process Z with mean offspring $1+\alpha/T$, where $T = T_a/\lambda$; T_a is the time to our most recent common ancestor, λ is the mean effective lifetime (or *generation time*) and $\alpha \in \mathbb{R}$ is our 'growth' parameter.

If we start time at the death of Eve then, in the notation of §3, $N_T(0) = 2$. (Eve, by definition, had at least 2 daughters with descendents alive today, and [13, Theorem 2.2] tells us that 3 such daughters is extremely unlikely: $N_T(0-) = 1$ and $N_T(0) \geq 2$ together imply that $N_T(0) = 2$ with high probability when T is large.) Note that $Z(T)$ is the current (effective) female population size.

Using our approximation results, we can simultaneously estimate α and T, based on the observations $Z(T)$ and the average pairwise divergence in a random sample of n contempory individuals $\overline{d}_n$. We will assume for the moment that the divergence rate δ is known. Denote by λ the mean effective lifetime of an individual. By [13, Theorem 2.1] (an exponential limit law for near-critical branching processes),

$$E(Z(T)|\ N_T(0) = 2) \simeq \frac{\sigma^2 T_a}{\lambda\alpha}(e^\alpha - 1). \tag{8}$$

We also have, by (7),

$$E\left(\overline{d}_n \middle|\ N_T(0) = 2\right) \simeq \delta T_a \gamma(\alpha), \tag{9}$$

where

$$\gamma(\alpha) = 1 - 2\alpha^{-1}\int_0^1 \frac{u}{(1-u)^3(u+\kappa(\alpha))}\left[1 - u^2 + 2u\log u\right] du, \tag{10}$$

and

$$\kappa(\alpha) = \frac{e^{-\alpha}}{1 - e^{-\alpha}}. \tag{11}$$

Note that $\gamma(\alpha)$ is positive and increasing in α, $1/3 < \gamma(\alpha) < 1$, and $\gamma(\alpha) \nearrow 1$ as $\alpha \to \infty$.

For the simplest moment based estimates, assuming that δ, σ^2 and λ are known, just set

$$Z(T) = \frac{\sigma^2 \hat{T}_a}{\lambda \hat{\alpha}}(e^{\hat{\alpha}} - 1), \tag{12}$$

$$\hat{T}_a = \frac{\bar{d}_n}{\delta \gamma(\hat{\alpha})}, \tag{13}$$

and solve for $(\hat{\alpha}, \hat{T})$. Although σ^2 is unknown, when α is sufficiently large the actual value (within reason) will not affect the estimates considerably. (This is due to the dominating exponential term in Equation 12.) The same is true for λ.

Note that in theory this approach assumes that α is small relative to T. However, since a large value of α corresponds to the (significantly) supercritical case, the estimate of T obtained from (13) will still make sense if the estimated value of α is large.

We would now like to apply our method to some data: but where does one find a *random* sample of individuals? Strictly speaking this is simply not available, as yet. However, we will do our best with what we have.

Of the 189 individuals considered by Vigilant *et al.* [22], we have hand-picked a somewhat representative sub-sample of 19, without being deliberately biased in any way. The larger the sub-sample, the less representative it becomes; the smaller it is, the less useful it becomes. Our sample consists of 6 Asians, 1 Native Australian, 1 Papa New Guinean, 6 Europeans and 5 Africans.

A histogram of the 171 pairwise divergences in this sample is shown in Figure 1. The average divergence was found to be 2.8%.

In June 1992, according to the *Population Reference Bureau Estimates*, the human population size was approximately 5.412 billion. This gives us about 1 billion as a rough estimate for the current effective female population size, assuming that about half the population is female, and that the current female population represents approximately 2.7 generations. We will soon

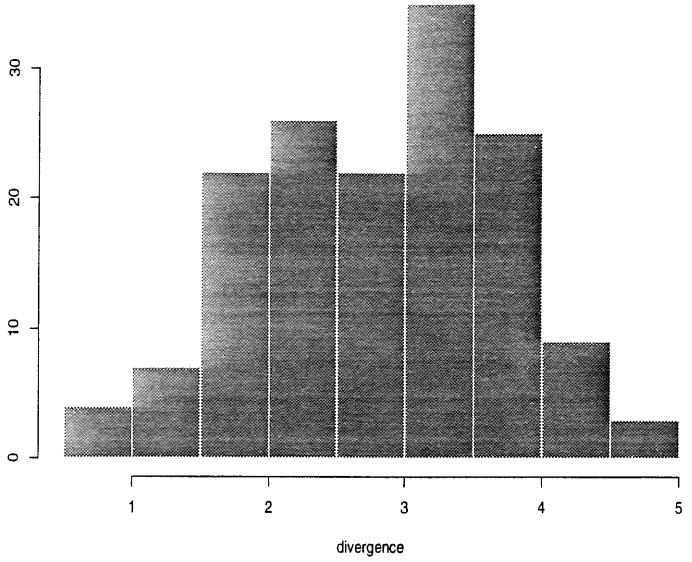

Figure 1: Pairwise divergences among sample of 19 individuals.

see that our estimates are quite insensitive to variations in this figure, so we needn't be very exact.

Note that the estimates $\hat{\alpha}$ and $\hat{T}_a$ are determined by $\lambda Z(T)/\sigma^2$ and δ; these are shown in Table 1, for various different values of $\lambda Z(T)/\sigma^2$ and δ. If $Z(T) = 1$ billion, $\sigma^2 = 2$ and $\lambda = 25$, then $\lambda Z(T)/\sigma^2 = 12.5$ billion. Although these choices seem somewhat arbitrary, we can see from Table 1 that any kind of realistic deviations from these values will have little or no effect on the estimates. The most important parameter is δ, the rate of divergence, and as yet there is no universally agreed 'best estimate' for δ. The values used in Table 1 are based on human-chimpanzee comparisons using a simple correction for multiple substitutions and possible dates of 9, 6 and 4 million years for human-chimpanzee divergence.

We conclude this section with some remarks on possible developments. To derive our estimates for the growth rate, α, and the age of Eve, T_a, we simply calculated the expected current population size and the expected average pairwise divergence in a sample of contemporary individuals, and assumed the other parameters were known. We are therefore not fully utilising the

Table 1: Estimates for α and T_a.

$\lambda Z(T)/\sigma^2$	δ	$\hat{\alpha}$	$\hat{T}_a$
12.5 billion	1.8	11.33	1,706,103
	2.7	11.77	1,113,265
	4	12.21	762,437
5 billion	1.8	10.31	1,722,531
	2.7	10.76	1,143,245
	4	11.2	768,607
30 billion	1.8	12.29	1,693,286
	2.7	12.94	1,125,364
	4	13.17	757,516

information contained in the sample. It might be helpful to know more about the joint distribution of the pairwise divergences (d_{ij}), or the joint distribution of the respective frequencies of distinct types, in a finite sample. The latter would be analogous to *Ewens' sampling formula* for the infinite-alleles Wright-Fisher model for neutral evolution. Ewens' sampling formula is not applicable to the Eve problem because it is based on the assumption that the population size is constant over time.

In particular, it may be possible to estimate α, T_a and δ simultaneously, without having to rely on human-chimpanzee comparisons, thus avoiding the assumption that the rate of divergence has been constant ever since the human and chimpanzee lines diverged.

Taib [19] has made some progress in this direction by obtaining an expression for the the asymptotic proportion of alleles (types) with exactly j representatives in the population, for a supercritical branching process with neutral mutations. Unfortunately, this result is not directly applicable here. The recent work of Pitman [15, 16] on sampling distributions may be very useful here.

5 Branching versus traditional models

The essential difference between branching and traditional models is that the latter assumes the total population size to follow some deterministic function over time, in which case the genealogy is described by a corresponding time-change of Kingman's coalescent. The relationship between the two is given by the fact that [14] if we condition a branching process on the evolution of its total population size, we have a traditional population genetics model.

For this reason, the qualitative behaviour of the two models is similar if the mean population growth rates are comparable. However, if a spatial component is introduced—for example, if individuals are allowed to immigrate between geographically separated colonies—the two models can exhibit radically different behaviour. This is because in a branching process, if the dynamics of migration and reproduction are independent, the spatial component does not influence the genealogy at all; in the corresponding traditional model (the general stepping stone model) where the population size in each colony is restricted to follow a deterministic course, the genealogy is strongly influenced by the degree of separation between colonies [2, 12].

A more realistic model would perhaps lie somewhere in between: population dynamics are certainly affected by geographical factors, but not to the extent which is assumed in traditional models. Although the details of such a model might be difficult to ascertain, a lot could be learned by studying the implied qualitative behaviour.

Acknowledgements. The author would like to thank all the participants of this workshop for many valuable discussions, and the organisers for making it happen.

References

[1] W. Bühler. The distribution of generations and other aspects of the family structure of branching processes. In *Proc. Sixth Berkeley Symp. Math. Statist. Prob., vol. III*, pages 463–480. University of California Press, Berkeley and Los Angeles, 1972.

[2] D. Dawson. Hierarchical and mean-field stepping stone models. Presented at this workshop.

[3] P. Donnelly and T.G. Kurtz. A countable representation for the Fleming-Viot measure-valued diffusion. Preprint, 1994.

[4] R. Durrett. The genealogy of critical branching processes. *Stoch. Proc. Appl.*, 8:101–116, 1978.

[5] J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.*, 27:401–410, 1978.

[6] M. Hasegawa and S. Horai. Time of the deepest root for polymorphism in human mitochondrial DNA. *J. Mol. Evol.*, 32(1):37–42, 1990.

[7] P. Jagers. The growth and stabilisation of populations. *Statist. Sci.*, 6:269–283, 1991.

[8] P. Jagers. *Branching Processes with Biological Applications.* Wiley, Chichester, 1975.

[9] P. Jagers, O. Nerman, and Z. Taib. When did Joe's great ... grandfather live? Or on the timescale of evolution. In: I.V. Basawa and R.L. Taylor (eds.), Selected Proceedings of the Sheffield Symposium on Applied Probability. IMS Lecture Notes–Monograph Series, Vol. 18, 1991.

[10] L. Krishtalka. *Dinosaur Plots and Other Intrigues in Natural History.* Avon Books, New York, 1989.

[11] R. Lundstrom, S. Tavaré, and R.H. Ward. Estimating substitution rates from molecular data using the coalescent. *Proc. Natl. Acad. Sci. USA*, 89:5961–5965, 1992.

[12] P. Marjoram. Population evolution and genealogy. Presented at this workshop.

[13] Neil O'Connell. The genealogy of branching processes and the age of our most recent common ancestor. To appear in *Adv. Appl. Prob.*, July 1995.

[14] E.A. Perkins. Conditional Dawson-Watanabe processes and Fleming-Viot processes. Preprint, 1992.

[15] J.W. Pitman. A two-parameter version of Ewens' sampling formula. 1992. Preprint.

[16] J.W. Pitman. Random discrete distributions invariant under size-biased permutation. *J. Appl. Prob.*, 1993. To appear.

[17] T. Shiga. A stochastic equation based on a Poisson system for a class of measure-valued diffusion processes. *Jour. Math. Kyoto. Univ.*, 30(2):245–279, 1990.

[18] Z. Taib. *Labelled Branching Processes with Applications to Neutral Evolution Theory.* PhD thesis, Chalmers University of Technology, Sweden, 1987.

[19] Z. Taib. The most frequent alleles in a branching processe with neutral mutations. 1991. Preprint.

[20] S. Tavaré and R. Griffiths. Estimation and inference in the coalescent. Presented at this workshop.

[21] A.G. Thorne and M.H. Wolpoff. The multiregional evolution of humans. *Scientific American*, April:76–83, 1992.

[22] L. Vigilant, R. Pennington, H. Harpending, T.D. Kocher, and A. Wilson. Mitochondrial DNA sequences in single hairs from a southern African population. *Proc. Natl. Acad. Sci. USA*, 86:9350–9354, 1989.

[23] L. Vigilant, L. Stoneking, H. Harpending, K. Hawkes, and A. Wilson. African populations and the evolution of human mitochondrial DNA. *Science*, 253:1503–1507, 1991.

[24] A.C. Wilson and R.L. Cann. Recent African genesis of humans. *Scientific American*, April:68–73, 1992.

[25] A. Zubkov. Limiting distributions of the distance to the closest common ancestor. *Theor. Prob. Appl.*, 20:602–612, 1975.